UNITED STATES DISTRICT COURT
DISTRICT OF MASSACHUSETTS

SCANSOFT, INC., Plaintiff, v. VOICE SIGNAL TECHNOLOGIES, INC., LAURENCE S. GILLICK, ROBERT S. ROTH, JONATHAN P. YAMRON, and MANFRED G. GRABHERR, Defendants.	)	C.A. No. 04-10353-PBS

DECLARATION OF CHARLES C. WOOTERS IN SUPPORT OF VOICE SIGNAL TECHNOLOGIES, INC.'S OPENING CLAIM CONSTRUCTION MEMORANDUM FOR U.S. PATENT 6,501,966

I, Charles C. Wooters, on oath, depose and say as follows:

1. I am a research scientist in the field of speech recognition, and am a Senior Research Engineer at the International Computer Science Institute, Berkeley, California. I hold a Ph.D. in Speech Recognition from the University of California, Berkeley. I also hold a Master's degree in Linguistics from the University of California, Berkeley. My curriculum vitae is attached as exhibit G to this Declaration.

2. The '966 Patent is entitled *Speech Recognition System For Electronic Switches In A Non-Wireline Communications Network*. As the title indicates, the '966 Patent describes an apparatus and method for providing speech recognition for a wireless (non-wireline) telephone network. The patent describes a particular user interface and an apparatus that is designed to enable users of a wireless network to connect to a requested phone number by voice. It does not

discuss (or claim) anything relating to *how* the speech recognition technology works – how particular sounds are recognized.

3. The user interface in a traditional telephone is the number pad. "Voice dialing" systems are designed to enable people to use voice commands instead of a keypad to call a particular number. The system described in the '966 patent has a user interface that provides a specific way for the user (once dialed into the system) to start the system and then to speak each digit of the desired number or to say a keyword that is associated with a telephone number that is stored in memory.

4. Systems that provide voice dialing capability have been available since the early 1980's and have been used in conjunction with land-line telephones (the phone on your desk), car phones (phones that are permanently installed in a car), cell phone handsets (today's mobile phone), CB radios, and in central network switch environments (Verizon's servers at the central switching office).

5. Voice dialing systems are made up of several distinct components which interact to provide the user with the desired functionality. The components are:

a. The user interface (UI) application. This is the software that receives the user commands and, based on those commands, controls the behavior of the system.

b. The speech recognition engine. This is the software that (a) decodes the user's speech; and (b) conveys back to the UI what has been said.

c. The hardware. This is the central processing unit (CPU) and memory that carry out all of the instructions and calculations directed by the software.

6. Any software system that interacts with a human being has what is called a user interface or "UI." This is the set of controls or steps presented to the user in a specific order to enable the user to control the functions of the system. On a more technical level, the UI interacts

with all of the components that make up a system, such that when the user selects a particular component, the UI instructs that component to perform its function. In the voice dialing application, the UI controls the prompts that are given to the user (e.g., "please say a keyword now"), the vocabulary that can be recognized at any given time, and the actions the system should take when a particular word is recognized. The user interface may also use prompts to provide feedback so that the user can verify the input before the system proceeds (e.g., "did you say 'home'?")

7. The speech recognition engine is a software program that decodes the sounds in human speech and compares them to speech templates to identify the words that have been spoken. In the context of the '966 patent, the speech recognition engine is treated as a black-box. The '966 patent does not disclose or claim any of the inner workings of the speech engine (called a "voice recognizer" or "speech recognizer" in the patent).

Speech Recognition Engines Generally Contain The Components Described Below.

8. The Front End. This is the module that extracts the unique "signature" elements (commonly called "features") of each sound. Many sounds have parts that are essentially the same (consider the sounds when the letters "C" "B" "D" "E" "G" "P" "T" are spoken). The job of the front-end is to throw away all of the redundant information that is not useful in identifying the sound, and pass the useful information to the decoder.

9. The Decoder. This is the module that compares the features received from the front-end to the previously stored templates for each sound. To perform this comparison, the sounds are broken into small units, known as phonemes (irreducible units of sound). The decoder compares the features it receives from the front-end to a library of phonemes previously

stored as templates. The output of the decoder is a sequence of phonemes that best matches the features received from the front-end.

10. The Templates (also called Acoustic Models). The templates used by the decoder are developed by feeding recordings of people saying the same words into a training system. The output is a model for each relevant sound. When the speech recognition engine is in operation, this model is used by the decoder to deduce the identity of the sound spoken by the user. The accuracy of a speech system depends heavily on the quality of the models, and the quality of the models depends heavily on the recordings used. Acoustic models can be built to be very "high resolution" in that they keep and represent the fine nuances of how sounds look, or they can be made to "low resolution" meaning they are gross approximations of sounds which throw away many of the details. The amount of memory used to store the models depends on the resolution selected by the developers of the speech recognition engine.

11. The Dictionary. This is a database of words with corresponding phoneme sequences. When the decoder produces the closest matching phoneme sequence, that sequence is looked up in the dictionary to identify the words associated with the phoneme sequence. The recognized words are then passed to the UI.

12. The Language Model. This is software that influences the output of the decoder so that only word sequences that make sense are produced. Since the decoder is not perfect, the language model helps it determine which phoneme (and therefore which word) to choose when the result is unclear. For example, if the decoder has determined that the last recognized word was "directions" and the next word has the sound "oo" and the choices are "to" or "you," the language model will influence the decoder to pick the sequence for "to" since the words "directions you" are less likely to make sense than "directions to."

13. Speech recognition software generally can be designed to be either speaker dependent or speaker independent. A voice recognizer might have both speaker-independent and speaker-dependent functionality.

14. Speaker Dependent (SD) speech recognition was developed prior to the early 1980's and has been widely available since that time. In general, speaker dependent speech recognition software requires the end user to first "train" the recognizer by speaking the exact word the user wishes the system to recognize. An advantage of speaker dependent systems is that the system's vocabulary can be tailored to, and modified by, the end user. For example, it can be constructed to recognize specific proper names.

15. Speaker Independent (SI) speech recognition systems were developed in the mid- to late 1980's. In general, they are systems that are "trained" in advance by the developer, rather than trained by individual end users. The developer "trains" a speaker independent system by feeding thousands of recordings of different people (with different accents, inflections and background noise) saying the same words into the system. These recordings are used to develop templates (also called "models") that are compiled into the system. Speaker independent systems have the advantage that the system can recognize most people's commands immediately "out of the box," with no need for individual user training. They also have the advantage that the recordings used for training can be made in a variety of different background environments so that the system is equally capable of later recognizing words spoken in a car, on a train or in a quiet office (for example). At the time of the '966 application, one disadvantage of speaker independent systems was that the recognizer's vocabulary was fixed and could not be modified by the end user. This limitation has since been overcome.

16. All other factors being equal, the larger the "active vocabulary" (this refers to the number of words that the speech engine can recognize at the same time), the greater the possibility of recognition error. One way to reduce the risk of error and still support a relatively large vocabulary is to divide the vocabulary into many discrete, but small active vocabularies, so that the system can recognize a relatively large number of words, just not all at once. The burden falls on the user interface to determine which sub-vocabulary needs to be active at any point in time. This can be achieved, for example, by providing a verbal prompt to a user, such as "please say a name" or "number please." For example, if the name "Juan" and the number "One" are in the same active vocabulary, the system could easily confuse the two, and might mistakenly recognize: "Six Juan Seven…", instead of "6-1-7…." If names are separated into one sub-vocabulary and numbers into another, the recognizer will not make this type of mistake.

17. Many technical approaches are known in the art for each of the above components in each variety of SD and SI. In fact there are thousands of published research papers describing each of the components of speech recognition engines described above and discussing particular implementations that offer various advantages and disadvantages.

Hardware and its influence on speech engines.

18. The impact that the hardware has on the design and implementation of a software system can not be understated. It is by far the most critical consideration any developer must analyze and design to when building a software system. It is critical to understand the limitations of the hardware that will be used to run the system *before* the software is written, so that the developer can make appropriate design decisions to enable proper operation. If, however, a software system is developed for one type of hardware and developers later attempt to modify it to run elsewhere, serious problems occur. For example imagine the difficulty of

getting Windows XP to run on the computer in your Microwave oven. Windows was not designed for that type of hardware, and so it would be difficult if not impossible to successfully port it (port means to adapt a software system to run on a hardware platform) to such a hardware platform.

19. Software is a very powerful tool in the hands of skillful developers, but once it is designed and implemented it becomes very inflexible. Software systems contain instructions to the CPU to retrieve data from memory, perform calculations, store intermediate results, move data back into storage, paint pixels on displays, interpret radio signals, decode and encode speech, and keep track of everything down to millionths of a second. Many software systems, such as speech recognition engines, perform calculations millions of times per second, and at each step many operations must be performed flawlessly. Common software systems use what are called "pointers" which direct the CPU to get a particular piece of data in a specific location in memory to perform an operation or calculation. When a software system contains, say, 200,000 lines of code and a developer mistakenly changes the location of a piece of data, the pointer will no longer work and the system will crash. If a speech system was developed for a PC and contains 200,000 lines of code, it is not trivial to shrink it to less than 50,000 lines to fit into a cell phone. The 150,000 lines that must be removed each had a purpose. One cannot simply strip them out of the system and expect it to function. If, however, the developer starts with the knowledge that the system can only be 50,000 lines, a functional system can be built by making appropriate trade-offs up front.

20. To recognize large vocabularies – for example the active vocabulary in a dictation system -- most speech engines require a significant amount of CPU power and memory. For example, the dictation system Dragon Naturally Speaking Version 8 requires 500 Megabytes of

hard disk space, 256 Megabytes of RAM, and a Pentium III 500 MHz CPU. Mobile phones today only have a tiny fraction of this CPU power and memory

Voice Dialing: Embedded Systems vs. Network Systems.

21. Cellular or mobile telecommunications systems connect mobile units to other telephone users and facilitate the transmission of those calls. Mobile units can take a number of different forms including car phones (phones permanently installed in cars) and mobile phones (today's handsets). In general a mobile unit is a small computer system with a radio transmitter and receiver for sending signals over the air to a radio tower.

22. When voice dialing software is added to a mobile unit it effectively replaces the keypad as a user interface. The voice recognizer is said to be "embedded" in the mobile unit. The user speaks into the unit, the sounds are passed to the speech recognition engine, interpreted by that engine and the result is passed directly to the mobile unit's conventional dialing mechanism. The number is then dialed just as if the keypad had been used.

23. When the voice dialing system is network based and located at the switch (MTX) certain technical challenges are presented. First, voice commands must travel *over the air* before reaching the voice dialing system. Anyone who has experienced a broken and garbled cell phone conversation knows that the voice can be distorted by the time it reaches you. This same problem occurs when dialing into a network based voice dialing system making the task of recognition much more difficult. The second challenge is to enable multiple users to access the same centrally located voice dialing system at once.

24. When the voice recognizer is network based, users must first manually dial into the voice recognition system by dialing the phone number for the network's speech recognition system (or perhaps pushing a speed dial button that dials the number). The network switch

establishes a connection between the user and the speech recognition system at the switch. The commands spoken by the user may then be transmitted to the voice recognizer over the wireless network in exactly the same way as in a normal person-to-person call. The centrally-located network based voice recognizer listens to the voice commands it receives over the wireless network and then, depending on which words were spoken by the user, will (in a voice dialing application) instruct the central switch to dial the requested number, connecting the mobile user to the desired recipient.

The Prior Art.

25. Certain prior art systems typify early voice dialing systems. During the processing of the application leading to the '966 patent, certain prior art relating to voice dialing was before the United States Patent Office. Some of that art described voice recognizers used in conventional (non-cellular) telephones. Other art described voice recognizers in cell phones.

a. U. S. Patent No. 4,348,550 to Pirz (Bell Labs) (Exhibit A, attached). The Pirz patent issued in 1982 from an application filed in 1980 -- twelve years before the filing of the application leading to the '966 patent. Pirz describes a voice dialing mechanism in a conventional telephone instrument. The system was speaker dependent. It allowed a user to "dial" a telephone number by speaking a series of digits or by speaking a person's name. The user first recorded certain "command" words -- *e.g.*, *off hook* -- as well as digits and names. Ex. A, 4:40-51. Thereafter, when the user spoke the command *off hook*, the voice recognizer "listened" for digits or a name. When the user spoke, and the recognizer recognized a digit, the recognizer assumed that a telephone number was being spoken, and thereafter compared incoming sounds only to its templates for digits. Upon recognition of the last expected spoken digit, the system said to the user, "I have recognized the number," followed by the digits that it

had recognized so as to allow the user to verify that recognition was accurate. Then, unless the user spoke the commands *stop* or *error*, the speech recognition system electronically transmitted the telephone number to the phone's dialer, which dialed the phone. Ex. A, 5:44-6:9. The same pattern was followed if a name was recognized, except that, upon recognition of a name, the recognizer had to "look up" the telephone number that was associated with that name in its memory. Ex. A, 6:9-42.

b. U. S. Patent No. 4,853,953 to Fujisaki (NEC Corporation) issued in 1989 on an application filed in 1988 (Exhibit B attached). This patent describes a speech recognizer in a telephone handset. The speech recognizer had both speaker-independent and speaker-dependent capabilities. A fixed set of "command words" and numbers were recognized by the speaker-independent functionality of the voice recognizer. The user trained the system's speaker dependent capability to recognize a group of names selected by the user and recorded a phone number to each name. Thereafter, the system could recognize each name selected by the user and would dial that number. Ex. B, 1:30-57, 2:64-3:3. The system repeated back to the user the words that it recognized to allow confirmation that it had interpreted the speaker's words accurately. *Id.*

c. In 1986, *Speech Technology*, an industry publication, contained an article written by Thomas B. Schalk, who was later named as an inventor in the '966 patent. (Exhibit C attached). The article described Mr. Schalk as the Director of Technology Development of Voice Control Systems. This article described a speaker-independent voice dialing system embedded in a cell phone. Ex. C, p. 24, col. 1. The article referenced the availability of speaker-dependent systems that could be trained to recognize names of persons and to allow associated phone numbers then to be dialed. Ex. C, p. 4, col. 2-3. The speaker-independent

system described in the article could recognize the numbers 0-9 (and "oh") and a group of words like *dial, speed dial, send, home, office, emergency* and the like. Ex. C, p. 26, col. 3. The system divided this vocabulary into sub-vocabularies. Its user interface required that a word selected from a particular sub-vocabulary be spoken first. Depending upon which word was spoken first, the user interface then activated other sub-vocabularies. Thus, for example, if the user first said *dial*, only the sub-vocabulary containing numbers would be activated. The user would then speak the digits of a phone number, which the voice recognition engine would match to the number vocabulary. When the user had finished, the user said *verify* and the system would repeat the numbers that it had recognized. If the number sequence was correct, the user was to speak the word *send* and the telephone number would be dialed. Alternatively, if the user first spoke the words *speed dial*, only the sub-vocabulary for speed dial names was activated. The user was then to speak a "destination description" like *home* or *office*." The recognized words were repeated to the user and, if recognition was correct, the system would obtain from memory the phone number that the user previously had associated with that destination description, and the user would say *send*. As noted above, the use of this type of "syntax" or "grammar" -- the speaking of a limited set of command words that, if recognized, inform the system of which sub-vocabulary to activate allows the speech recognition task to be more accurate.

d. Substantially the same product was described by R. Eugene Helms, Director - Product Development of Voice Control Systems, in a paper presented at Speech Tech, an industry conference, in 1986. (Exhibit D attached.) The Helms paper explained that:

> By using a voice command syntax structure, similar to a command menu structure common in man-computer interfaces, large command vocabularies can be accommodated in voice control applications by partitioning the total vocabulary into a set of smaller vocabularies. Each such sub-vocabulary would then correspond to nodes in the syntax tree. This approach is appealing because it retains simplicity for the user

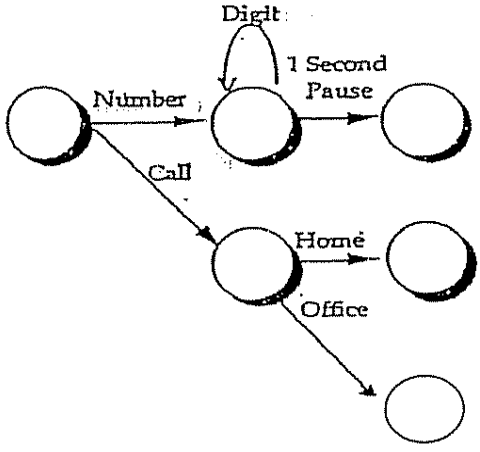
> through a logical structuring of the commands while effectively making the voice recognition task one of recognizing a number of smaller vocabularies rather than one large vocabulary. The latter point has strong implications for the accuracy of the voice recognizer . . . .

The Helms paper noted the existence of speaker-dependent and speaker-independent systems. It said speaker-dependent systems were simpler and more flexible than speaker-independent systems, but had the disadvantage that users were required to "train" the voice recognizer to recognize the user's speech. The article then described Voice Control Systems' speaker-independent product, again explaining the "syntax" used by the voice recognizer and the way in which the system verified that its recognition was accurate.

> As indicated in Figure 4 above, five primary commands may be given to the phone. They are DIAL, SPEED-DIAL, RECALL, EMERGENCY, and SEND. The DIAL command is followed by a sequence of discrete digits which is terminated by the command VERIFY, whereupon the digit sequence is repeated to the user for validation. If the user wants the call to then be placed, he says SEND; otherwise he can issue the command CLEAR to effectively restart the command sequence. The command SPEED-DIAL is followed by one of a set of call destination descriptors such as HOME, OFFICE, SCHOOL, etc. Each of these descriptors is associated with a unique memory location. When the word HOME is recognized, for example, the phone number previously stored in memory location N is accessed, whereupon the voice response unit says "DIALING HOME". Again, if the user desires to place that call, he utters the command SEND; otherwise he can say CLEAR to abort the call placement. The command RECALL is functionally the same as SPEED-DIAL except that a two-digit memory location is entered directly following the command. This allows repertory dialing using numeric speed codes. The primary command EMERGENCY will immediately effect call placement to the emergency number in memory without qualifying commands or validation.

Ex. D, p. 129-130

e. The voice recognition system embedded in the mobile unit described in the Schalk and Helms articles was incorporated in a product called VoiceDial offered by Uniden. (Exhibit E attached). The Uniden Operating Guide, which was before the Patent Office, states


Digit
1 Second
Pause
Number
Call
Home
Office

that "using your cellular phone with VoiceDial is not only easy, it adds safety, speed and convenience." Ex. E, p. 1. The Guide then describes the syntax, prompts and verification technique used to voice dial a telephone number. Ex. E, p. 4. The user says *phone . . . start*. The phone responds "ready." The user says *dial*. The phone says "number, please." The user speaks the phone number that the user wishes to dial and then says *end*. The phone repeats to the user the phone number that it has recognized. If the phone number was correctly recognized, the user says *send*. If the user wishes to call *home, office* or any one of several other locations, the user begins with the command *call* (instead of *dial*), and a similar series of prompts, responses and recognition verification steps followed. VoiceDial also included a speed-dial functionality that allowed a user to speak a two-digit number (*e.g.*, 17) in place of a destination description (*e.g.*, *home*).

f. A similar voice dialing product embedded on a mobile unit is described in the publication *Machine Design* in January, 1991. (Exhibit F attached). The authors were Savaraj Pawate and Peter Ehlig of Texas Instruments. The system used what it called a "grammar" similar in structure and purpose to the "syntax" described in the Voice Control Systems/Uniden system described above. However, unlike that system, the Texas Instruments' product could recognize (and cause cell phone calls to be placed to) a spoken telephone number, a spoken location (*e.g.*, *"office"*) and to a spoken "repertory name, for example, 'call Harvey.'" Ex. F, p. 96. A sample of the grammar used by this product is depicted in the flow chart on the facing page.[1]

[1] The article states that "Other application grammars are possible. An application may, for example, require that the speech recognition system recognize names and the word *call* as in the command *call John Jones*."

The voice dialer used both ROM (read only memory) and RAM (random access memory). ROM supported a fixed-vocabulary, speaker-independent voice recognition capability. RAM would support a user-specific, speaker-dependent functionality. Ex. F, p. 96.

26. Each of the patents and publications described above except the Pirz reference were all before the Patent Office and are cited in the '966 patent. All describe voice dialing functionalities contained either in landline handsets or mobile cellular units. Presumably, they provided a backdrop for what could, and could not, be claimed in the '966 patent.

27. The '966 patent is entitled *Speech Recognition System For Electronic Switches In A Non-Wireline Communications Network*. As the title indicates, the '966 Patent describes a voice dialing apparatus and method where the voice recognition activity occurs at, or in conjunction with, the central switching apparatus of a cellular or other non-wireline network. The patent describes a centrally located voice recognition function that is shared by multiple users of a non-wireline network as well as the spoken grammar or syntax used by the system's user interface to direct a telephone call. It does not discuss the way in which the system recognizes particular words.

28. The patent begins with a Brief Summary of the Invention. The first portion of the Summary explains that the object of the "invention" is to provide a speech recognition system "for use at a mobile telephone exchange" of a cellular or personal communications network. 1:49. The patent states (1:44-65):

> It is therefore an object of the present invention to describe an implementation of a speech recognition system in a cellular or personal communications network environment.
>
> It is a further object of the invention to describe a speech recognition system for use *at* a mobile telephone exchange (MTX) of a cellular or personal communications network. The placement of the speech recognition system *at* the MTX significantly reduces cost and increases

> reliability by enabling the switch vendor to install and maintain the system *in conjunction with* the cellular switch.
>
> It is another object of the invention to describe a cellular voice dialing system for use *in or in conjunction with* an MTX of a cellular network.
>
> * * *
>
> Another object of the invention is to provide for combined use of speaker-dependent and speaker-independent voice recognition and speaker verification techniques *in* an MTX of a cellular or personal communications telephone network.

(Emphasis added)

29. Having described the object of the invention, the patent then explains how that object is achieved. Spoken commands are to be transmitted over the network to a centrally-located voice recognition system. Voice recognition is to be a shared resource -- a resource available to all users of the network -- and is to have both speaker-independent and speaker-dependent functionality.

The patent states (1:66 - 2:12):

> These and other objects of the invention are provided in an advanced system for the recognizing of spoken commands *over* the cellular telephone or any personal communications (*i.e.*, any non-wireline) network. In the cellular application, for example, a Speech Recognition System *interconnects either internally with or as an external peripheral to a cellular telecommunications MTX switch*. The Speech Recognition System includes an administrative subsystem, a call processing subsystem, a speaker-dependent recognition subsystem, a speaker-independent recognition subsystem, and a data storage subsystem. The Speech Recognition System also allows for increased efficiency in the cellular telephone network *by integrating with the switch or switches as a shared resource*.

(Emphasis added)

30. The patent then proceeds to a section entitled "Detailed Description." It makes clear that the invention is a "switch-based Speech Recognition System." When it adds

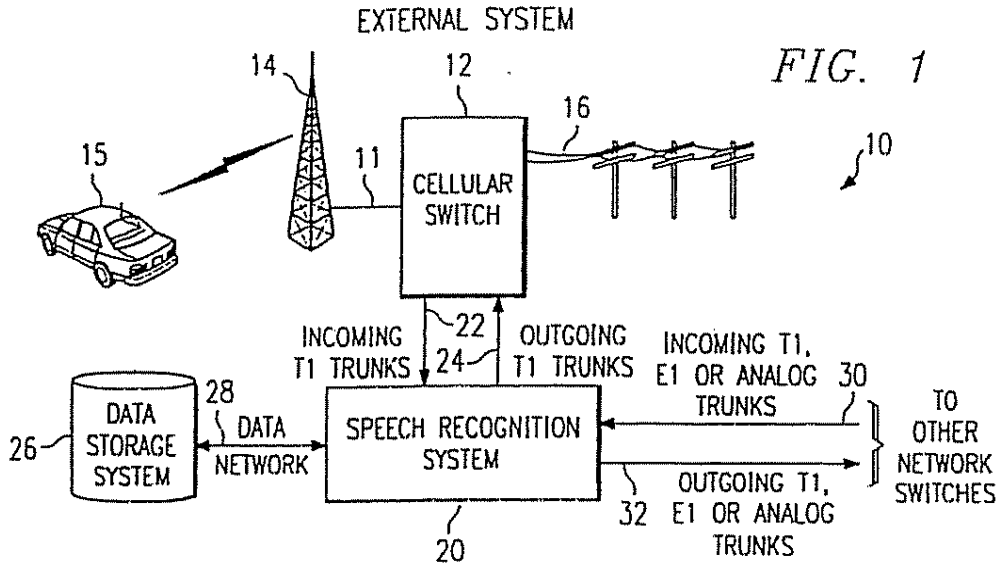
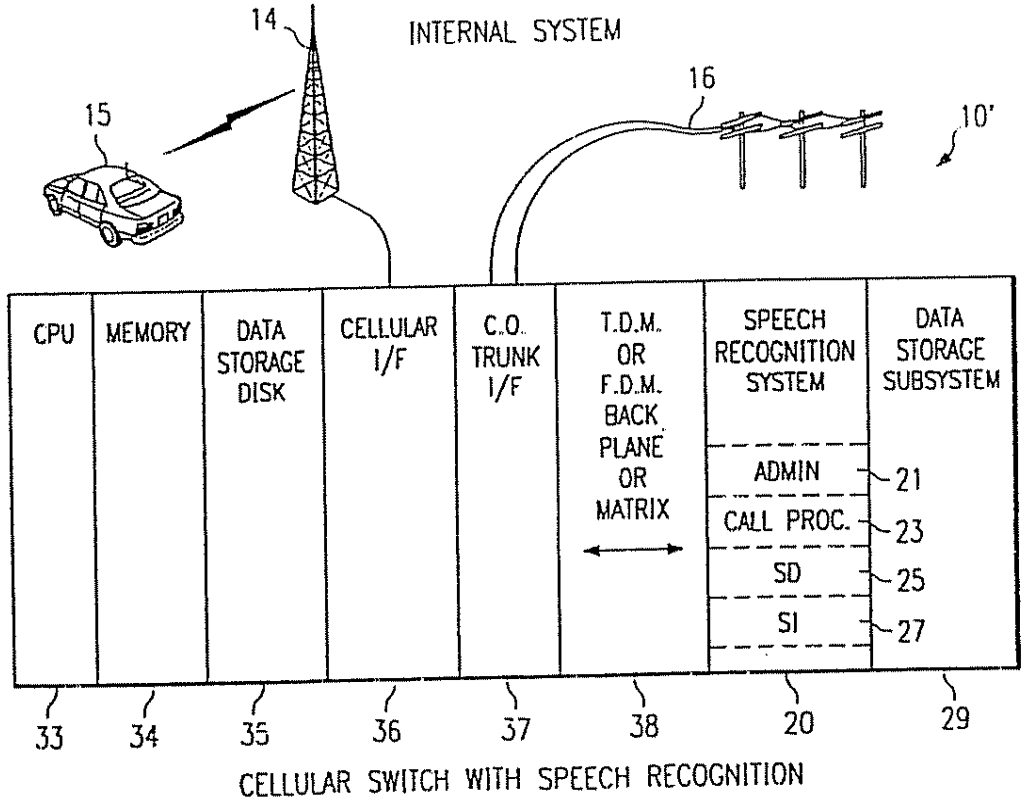


FIG. 1

20 TRUNKS

FIG. 2



CELLULAR SWITCH WITH SPEECH RECOGNITION

broadening language, it points out that it is referring to any non-wireline network, not merely a cellular network. It does not suggest that the "invention" includes speech recognition system embedded in a mobile unit -- *i.e.*, a non-shared resource. The Detailed Description begins with an explanation of Figure 1 of the '966 patent, which is reproduced in the facing page. It states (3:10-21):

> Fig. 1 is a block diagram of a cellular telephone network incorporating an external switch-based Speech Recognition System according to the present invention. Although the following description is specifically related to use of the Speech Recognition System, at or in conjunction with an MTX of a cellular network, it should be appreciated that the System also interconnects either internally with or as an external peripheral to a personal communications (*i.e.*, any non-wireline) network.

31. Note that in Figure 1, the Speech Recognition System (20) is shown as physically connected to the "cellular switch" by trunk lines (22 and 24). The patent explains (3:66-4):

> According to one embodiment of the invention as shown in FIG. 1, a Speech Recognition System (20) is connected as an external peripheral to the MTX through a set of preferably digital trunk lines. Set 22 is used for incoming signals and set 24 is used for outgoing signals. . . . The Speech Recognition System (20) may be integrated with one or more switches (whether or not cellular) for use as a shared resource via incoming and outgoing trunk sets 30 and 32.

32. The patent then describes a second, and the preferred, embodiment. In this embodiment, the speech recognition functionality is connected internally to the mobile telecommunication exchange (the central switch). This embodiment is depicted in Figure 2 (shown on the facing page) and is described in the patent as follows (4:13-26):

> Referring now to FIG. 2, a block diagram shows the cellular telephone network (10) with the Speech Recognition System (20) interconnected internally to the MTX. This is the preferred embodiment of the invention. The hierarchical architecture of the cellular switch includes the central processing unit (33), memory (34), data storage disk (35), cellular interface (36), central office trunk interface (37) and a backplane or switching matrix (38).

> The Speech Recognition System 20 includes a number of functional subsystems: an administrative subsystem 21, a call processing subsystem 23, a speaker-dependant recognition subsystem 25, a speaker-independent recognition subsystem 27, and the data storage subsystem 29 (which corresponds to the storage system 26 of FIG. 1)

33. No other embodiment is described. In particular, it is not suggested that the "invention" include speech recognition software embedded in a mobile phone. To the contrary, claim 1 of the '966 patent requires a "speech recognition method for a mobile telecommunications system . . . ." The body of the patent defines the term *mobile telecommunications system* as that which connects wireless customers to land-line customers (3:35-47):

> A cellular mobile telecommunications system connects mobile telecommunications customers, each having a mobile unit, to land-based customers served by a telephone network. . . . As used herein, "mobile telecommunications systems" refers to cellular, satellite and personal communications network environments.

Note the distinction between the "mobile telecommunications system" and the term "mobile unit" (elsewhere referenced in the patent as a "cellular telephone", *e.g.*, 3:49, 5:63).

34. The patent then proceeds to a description of how the "invention" works.[2] The method begins when a cell phone user dials into the speech recognition functionality and the switch connects the user to the speech recognition function. (5:59-6:3). *See* Fig. 5 at 102 ("User dials digits from cellular phone"). The dialing activity at the cell phone is entirely conventional. Unlike the prior art cell-phone-based systems, the speech recognition engine is not embedded in the mobile unit and is not an interface between the user and the dialing mechanism in that user's

[2] I have not addressed the hardware that is described in the '966 patent because specific hardware does not appear to be claimed in the '966 patent. I observe, however, that the invention is said to include a mass storage hard drive device (49). (4:48). *See* Fig. 3. The inclusion of such a device is wholly inconsistent with a cell-phone-based (as distinguished from a network-based) speech recognition system.

phone. Instead, the user communicates with the speech recognition functionality over the wireless network. The patent states:

> Referring now to FIG. 5 [see the facing page], a functional flowchart is provided describing the basic control functions of the Speech Recognition System according to the present invention. The routine begins at step 102 when the user dials digits from the cellular telephone. At step 104, a test is performed to determine if a Speech Recognition System access code [In effect, the telephone number of the speech recognition system] has been dialed. If not, the cellular switch processes the call based on the number dialed at step 106 and the routine ends. If the result of the test at step 104 is positive, the routine continues at step 108 during which the switch makes an audio path connection between the user and the Speech Recognition System.

35. The patent explains that each cell phone instrument is uniquely identified by a Mobile Identification Number or "MIN", which is nothing more than the telephone number of that telephone instrument. (3:49-53). When the switch connects the cell phone user to the speech recognition functionality, it also transmits the phone's MIN to the speech recognition system. The system then uses the MIN to fetch from the system's mass storage device the data that are unique to that particular user. Those data include the particular telephone numbers that the user has associated with a generic location identifier like *home* or *office*. This activity is required in a systemwide voice recognition functionality because different users have different phone numbers associated with generic (and speaker-independent) words like *home* or *office*. The MIN is also used to retrieve from mass storage proper names that have been input by that user into the system's speaker-dependent functionality as well as the phone numbers associated with those names. This allows the system-wide speech recognition functionality to search the limited roster of stored names used by a particular user, rather than searching through all names input by all users. The patent states (6:3-15):

> At step 110, the switch sends the user's mobile identification number ("MIN") to the Speech Recognition System. As noted above, the MIN is a

> unique number associated with a given cellular telephone that is available to the switch each time a telephone call is placed.
>
> According to the invention, each user who subscribes to the service will have prerecorded a list of destination numbers. At step 110, these speed-dial numbers, along with speaker-dependent templates [*i.e.*, proper names] and user language type data [indicating the language in which prompts are to be spoken to that user and commands given by that user], are retrieved from the data storage subsystem. As noted above, the data storage subsystem stores such data at predetermined locations that are preferably accessed by the MIN.

36. The patent then describes a syntax or grammar of the type described in the Voice Control Systems/Uniden product and Texas Instruments article referenced above. The system also uses a series of prompts, as described in the prior art, that are designed to cause the speaker to speak appropriate words in a particular order. The patent thus describes the first tier of the syntax to be used, and explains that depending upon which command is spoken, different user-interface routines are initiated and different sub-vocabularies are activated. It explains (6:15-35):

> The routine then continues at step 112 with the Speech Recognition System prompting the user that it is "Ready For Command" or the like. The command is made in the language as determined by the user language type data retrieved at step 110. At step 114, the Speech Recognition System engages the speaker-independent recognition subsystem to obtain the user response. Depending on the response, one of several different subroutines follow.
>
> If the user states and the system recognizes a "Dial" command, control is passed to the routine of FIG. 6. In particular, a test is made at step 116 to determine if the Dial command is recognized. If so, control is transferred to the routine of FIG. 6 [describing the speaking of a telephone number]. If the response to the test at step 116 is negative, a test is made at step 118 to determine if a "Call" command has been spoken and recognized. If the system recognizes a "Call" command, control is passed to the Call Routine of FIG. 7 [describing the speaking of a destination descriptor like "home" or a speed dial number]. If the response to the test at step 118 is negative, a test is made at step 120 to determine if a "Directory" command has been spoken and recognized. If the system recognizes a "Directory" command, control is passed to the Directory Dialing Routine of FIG. 8 [which describes the speaking of a proper name].

37. If the command "dial" is spoken, and if that command is recognized, the user-interface causes the system to say to the user "phone number, please." The user then speaks, and the system then collects the digits that make up a phone number. *See* 6:47-7 and Fig. 6.[3] If the phone number is recognized accurately, the MIN and the string of digits is passed to the switch (not a cell phone's dialer) and the switch then dials the telephone number and connects the user to the intended recipient of the call. 7:28-32.

38. If the command *call* is spoken, and recognized, the user may either speak a two-digit speed-dial number or what the patent calls a keyword, like *home, office* or *information*. If recognized, the system retrieves the phone number that the user has previously associated with the speed-dial number or keyword (8:9) and transfers the user's MIN and the relevant phone number to the network switch for dialing. The switch then dials the number. 8:19-25.

39. If the command *Directory* is spoken, and recognized, the system prompts the user to say a name and the system engages the speaker dependant recognition subsystem. A name is spoken. If the system recognizes the name, it fetches the phone number that the user has previously associated with that name, and transfers that phone number to the switch for outdialing. 8:26-41.

40. The patent then reiterates that the *invention* is a network-based voice recognition system. It distinguishes the prior art by stating that the invention can deal with many calls at once and states that algorithms have been developed that overcome the distortion of speech that frequently occurs when spoken words travel over a cellular network to a centrally-located speech recognition function. (11:21-61).

---

[3] Referring now to FIG. 6, the Dial Routine is described in detail. At step 130, the Speech Recognition System prompts the user with a message, such as "Phone Number, Please," and applies the speaker-independent recognizer to collect the digits.

> The present invention has numerous advantages over the prior art. The system combines the use of both speaker-dependant and speaker-independent speech recognition in an [sic] mobile or portable telephone communications network environment. Multiple language prompts are spoken from and available simultaneously on multiple ports from a single automated telecommunications-based system. The language selected is based on the language spoken by the user....
>
> The invention successfully implements speech recognition in the cellular telephone or personal communications network. Non-wireline networks provide a special challenge to both the recognition algorithm developers as well as the applications developers. The recognition algorithm in conjunction with the system application is insensitive to the radio fading, speech clipping, and speech compression conditions that occur in a non-wireline network. In addition, the recognition algorithm accommodates conditions found in the standard switched network. The invention provides a means of accurately recognizing speech that has limited distortion due to clipping or fading and provides a means of reprompting the user for input when the speech has become too distorted for accurate recognition.
>
> Previously, only the best examples of spoken words have been used as tokens for developing speech vocabularies. By collecting speech that has been compressed or that although distorted by radio fading or clipping is still intelligible and by adding this collected speech to the speech training database, the vocabulary based on such data becomes more robust and less sensitive to these conditions. Adding the distorted but intelligible data to the training database of excellent example words allows for a more diverse statistical representation of each vocabulary word. Words that might have been previously rejected because part of the word was clipped can now be recognized if enough intelligible information is available. If the statistical representation of the word indicates that not enough information is available for accurate recognition, the recognition system will reject the word and reprompt the user for input.

None of this would be relevant if voice recognition functionality were embedded in a single mobile phone.

41. Claim 1 of the '966 patent is set forth below:

> A speech recognition method for a mobile telecommunication system which includes a voice recognizer capable of recognizing commands and characters received from a mobile telecommunication user, the method comprising the steps of:

> receiving a command from the mobile telecommunication user;
>
> determining whether the command is a first or second command type;
>
> if the command is the first command type, collecting digits representing a telephone number to be dialed received from the mobile telecommunication user; and
>
> if the command is the second command type, determining whether a previously stored telephone number is associated with a keyword received from the mobile telecommunication user.

42. I understand that the parties disagree as to the meaning of the phrase "A speech recognition method *for a mobile communication system . . . .*" As noted above, the phrase *mobile telecommunication system* is defined in the patent as that which "connects mobile telecommunications customers, each having a mobile unit, to land-based customers served by a telephone network." The voice recognition method is "for" such a system. This definition does not treat a single "mobile unit" (cell phone) as a mobile telecommunications system or as part of such a system. I find in the '966 patent no suggestion that the invention of the patent includes a voice recognition method used by a single cell phone, and I do not believe that a voice recognition method that is entirely internal to a single cell phone is described anywhere in the patent.

43. Also, the fact that this is a voice recognition method *for a mobile telecommunication system* (as distinguished from a cell phone) is what makes the system of the patent different from earlier mobile phone-based speech dialing systems that were cited to the Patent Office as prior art. The steps required by the patent (receiving a command, determining whether it is a first or second command type, and so on) are indistinguishable from the syntax described in the prior art Voice Control Systems/Uniden product or the Texas Instrument article, as are the tasks of collecting digits representing a phone number and determining whether a

phone number is associated with a keyword spoken by a mobile telecommunication user. Furthermore, in looking at the steps of Claim 1, I note that in an embedded system, all of the steps of the claim are complete before the mobile unit begins its dialing interaction with the mobile telecommunications network. In fact, all of the steps of claim 1 could be completed without any interaction with a network telecommunications system.

44. I will comment briefly on two other claim construction issues. First, the series of steps stated in Claim 1 appears to require (1) the receipt of a command (e.g. "dial" or "call"), (2) a determination as to which command was spoken (e.g. "dial" or "call") and thus whether digits are to follow or a keyword is to follow, (3) if it is determined that digits are to follow -- *i.e.*, if a first command type (e.g. "dial") has been spoken and recognized-- digits representing a phone number are collected, and (4) if it is determined that a keyword is to be spoken -- *i.e.*, if a second command type ("call") has been spoken and recognized -- a keyword is received and the system then determines whether a previously-stored telephone number is associated with that keyword. This interpretation is consistent with the separate steps described in the '966 patent.

45. It is also consistent with the words used in the patent. The patent references the words *spoken in the first tier of the syntax -- i.e., Dial, Call* and *Directory* -- as "commands." *See* 6:24, 6:27, 6:29, 6:30, 6:33, 6:34. The '966 patent calls the numbers in a telephone number "digits" (not "commands") (6:50) and refers to other words that are designed to substitute for a telephone number "keywords" or "names" (not "commands"). 7:60-67, 8:26-34.

46. I understand ScanSoft states that the phrase "*dial 617-248-5207*" is a single command. This interpretation is inconsistent with the words of the claim which require a voice recognizer to recognize "commands and characters." The word *characters* appears to refer to the digits in a phone number, thus distinguishing those digits from "commands." Similarly, the

claim refers separately to commands and keywords. Therefore, the more natural interpretation of the phrase "*call home*" is that *call* is the command and *home* is the keyword.

47 Finally, I understand that ScanSoft states that the phrase *collecting digits representing a telephone number* requires the system to have intelligence that permits it to know that the user has spoken a predetermined number of digits. This would appear to add to the claim a requirement that is not included in the claim language. If a user speaks and the recognizer recognizes a series of digits that are, in fact, a telephone number, the system has "collect[ed] digits representing a telephone number." There are many known ways for a system to determine that it has received the last digit in a string. The claim does not appear to distinguish between systems that "expect" to hear a defined number of digits (*see, e.g.*, Ex. A at 5:48), or learn that the proper number of digits has been spoken because the speaker stops speaking for a defined time period (*see, e.g.*, Ex. F at 97), or the speaker cues the system that a full phone number has been spoken by saying a word (*see, e.g.* the Voice Control Systems/Uniden product). All would seem to be encompassed within the phrase *collecting digits representing a phone number*.

Sworn to under the pains and penalties of perjury this 5th day of June, 2005.

Charles C. Wooters